

NAVNING GENETIK TOZALIGINI TAHLIL QILISH UCHUN TAHRIR MASOFALARIDAN FOYDALANISH

Mirzabayev Dmitriy Alekseyevich, Abduraximov Azizjon Azim o'g'li, Radjapov

Farxod Sunnatullayevich, Salaxutdinov Ilxom Baxtiyarovich

O'zbekiston Respublikasi Fanlar Akademiyasi genomika va bioinformatika markazi

E-mail: dmitriy08m@gmail.com

Annotatsiya: Genetik tozalik bu - urug'lik materialining asosiy sifat ko'rsatkichlarini, shu bilan bir qatorda taqdim etilayotgan nav namunasiga qanchalik muvofiqligini belgilovchi mezon hisoblanadi. Hozirgi vaqtda genetik tozalikni aniqlash tajribalarida molekulyar SSR-markerlar yetakchi o'rin egallaydi. Shu nuqtayi nazardan, markerlar yordamida olingan ma'lumotlar allel mavjudligi (1) yoki uning yo'qligi (0) binar formatda tahlil etiladi. Bunda zaruriy shart, tahlil uchun mos o'lchov usuli (funktsiya) ni tanlash hisoblanadi. Ushbu maqolada SSR markerlarining binar ma'lumotlaridan foydalanib, genetik tozalik qiymatini hisoblash uchun Xemming va Levenshteyn oralig'i usullari asosidagi tahlillar keltirilgan.

Kalit so'zlar: genetik tozalik, Xemming oralig'i, Levenshteyn oralig'i.

ИСПОЛЬЗОВАНИЕ РАССТОЯНИЙ РЕДАКТИРОВАНИЯ ДЛЯ АНАЛИЗА ГЕНЕТИЧЕСКОЙ ЧИСТОТЫ СОРТА

Мирзабаев Дмитрий Алексеевич, Абдурахимов Азизжон Азим Угли, Раджапов

Фарход Суннатуллаевич, Салахутдинов Ильхом Бахтиярович

Центр геномики и биоинформатики Академии наук Республики Узбекистан

E-mail: dmitriy08m@gmail.com

Аннотация: Генетическая чистота является главным критерием для определения качества семян, а также степени соответствия представленного образца семян заявленному сорту. В настоящее время при определении генетической чистоты, ведущая роль отводится молекулярным SSR – маркерам. В этом отношении, данные, полученные с помощью маркеров, рассматриваются в бинарном формате присутствия (1) или отсутствия (0) аллеля. При этом необходимым условием является выбор подходящей метрики (функции). В

настоящей работе представлен анализ с использованием расстояний Хемминга и Левенштейна для расчета генетической чистоты с использованием бинарных данных SSR-маркеров.

Ключевые слова: генетическая чистота, расстояние Хемминга, расстояние Левенштейна.

USE OF EDITING DISTANCES FOR ANALYSIS OF GENETIC PURITY OF A VARIETY

Dmitry Mirzabayev, Azizzhon Abdurakhimov, Farkhod Radzhapov, Ilkhom
Salakhutdinov

Center of Genomics and Bioinformatics Academy of Sciences of the Republic of
Uzbekistan

E-mail: dmitriy08m@gmail.com

Abstract: Genetic purity is a key criterion for determining seed quality and the degree of conformity of the submitted seed samples to the declared cultivar. Currently, SSR molecular markers play a leading role in assessing the genetic purity of plant varieties. In this context, data obtained using these markers are interpreted in a binary format indicating the presence (1) or absence (0) of an allele. Accordingly, the selection of an appropriate metric (function) is essential. This paper presents an analysis of Hamming and Levenshtein distances for calculating genetic purity based on binary SSR marker data.

Keywords: genetic purity, Hamming distance, Levenshtein distance.

KIRISH.

Nav genetik tozaligi tahlili urug'lik partiyasi bo'yicha uning tozaligini baholashda hal qiluvchi mezon hisoblanadi. Genetik tozalikni tahlil qilish usullaridan biri Grow-Out Testdan hisoblanadi [1]. Biroq, ushbu usul morfologik belgilarni baholashga asoslanganligi sababli, ularning atrof-muhitga yuqori sezgirliги ushbu yondashuvni qo'llashni cheklaydi. Biroq, ushbu usul morfologik belgilarni baholashga asoslanganligi hamda ularning ushbu belgilarning atrof-muhit ta'siriga yuqori sezgirliги bu yondashuvni qo'llashni cheklaydi. Shu bilan bir qatorda muqobil usul sifatida, genotip va atrof-muhitning o'zaro ta'siriga bog'liq bo'lmagan

molekulyar DNK markerlaridan (SSR -Simple Sequence Repeats) foydalanish taklif etiladi [1, 2]. SSR markerlar yordamida olingan ma'lumotlar allel mavjudligi (1) yoki uning yo'qligi (0) binary pattern ko'rinishida ifodalanishi mumkin. O'z navbatida, genetik tozalikni sinovdan o'tkazish allel profillarini tegishli o'lchovlar - ular o'rtasidagi o'xshashlik yoki farq darajasini baholovchi funksiyalar yordamida taqqoslash orqali amalga oshirilishi mumkin. Ushbu tadqiqot ishida SSR markerlar binar ma'lumotlaridan foydalangan holda navning genetik tozaligini aniqlash uchun tahrirlash (Xemming va Levenshteyn oralig'i) asosida masofa o'lchov usullari ko'rib chiqiladi.

Xemming oralig'i bir xil uzunlikdagi ikki qatorning belgilari mos kelmaydigan pozitsiyalar sonini o'lchaydi. Binar patternlar (1, 0) yordamida allel profili taqqoslanganda, Xemming oralig'i quyidagi matematik xossalar bo'yicha bir qatorni boshqasiga o'zgartirish uchun kerak bo'lgan belgilarni almashtirishning minimal sonini hisoblash orqali aniqlash mumkin [3]:

$$d_H(x, y) = 0 \Leftrightarrow x = y$$

$$d_H(x, y) = d_H(y, x) > 0 \Rightarrow x \neq y \quad (1)$$

$$d_H(x, z) \leq d_H(x, y) + d_H(y, z)$$

1 – xossa bo'yicha allel mavjudligi/yo'qligining binar patternlarini tahlil qilish uchun beshta namunani (A_1 - A_5) o'z ichiga olgan tanlanmada beshta SSR-markerlaridan foydalangan holda «A» navining genetik tozaligini baholashni amalga oshiramiz. Allelning mavjudligini 1 va uning yo'qligini esa mos ravishda 0 soni bilan bilan belgilaymiz (1 – jadval).

1-jadval.

«A» navining allel profili

Namuna	SSR – marker				
	1	2	3	4	5
A_1	1	1	0	1	1
A_2	1	0	1	1	0
A_3	0	1	1	0	1

A₄	1	0	1	1	1
A₅	1	1	0	1	1

1 – xossaga ko'ra, namunalar allel profili ma'lumotlari bo'yicha bir-biri bilan taqqoslaganda quyidagi d_H taqsimotini olamiz (2-jadval, a). Keyinchalik, d_H metrikalarining olingan qiymatlarini satr uzunligi (n) - d_H/n bo'yicha normallashtirish maqsadga muvofiqdir [4]. Demak, har bir namuna bir-biri bilan taqqoslanganda normallashtirilgan d_H quyidagicha ko'rinishda bo'ladi (2-jadval, b).

2-jadval.

«A» nomli nav namunalari solishtirilganda d_H ning taqsimlanishi

a)	A₁	A₂	A₃	A₄	A₅	b)	A₁	A₂	A₃	A₄	A₅
A₁	0	3	3	2	0	A₁	0,0	0,6	0,6	0,4	0,0
A₂	3	0	4	1	3	A₂	0,6	0,0	0,8	0,2	0,6
A₃	3	4	0	3	3	A₃	0,6	0,8	0,0	0,6	0,6
A₄	2	1	3	0	2	A₄	0,4	0,2	0,6	0,0	0,4
A₅	0	3	3	2	0	A₅	0,0	0,6	0,6	0,4	0,0

Tahlil qilinayotgan guruh («A» navi) ichidagi namunalar o'rtasidagi genetik tozalikning o'rtacha qiymatini baholash uchun barcha noyob juft taqqoslash namunalari ($A_1 - A_5$) uchun normallashtirilgan d_H ning o'rtacha qiymatini hisoblaymiz, bunda noyob juftlar soni (k) quyidagi ko'rinishda bo'ladi:

$$k = \frac{N \times (N - 1)}{2} \quad (2)$$

bu yerda,

N – namunalar soni.

2-formulaga ko'ra, 5 ta namuna ($A_1 - A_5$) uchun noyob juftliklar soni quyidagicha hisoblanadi:

$$k = (5 - 1) \times 5 / 2 = 10$$

O'z navbatida, namunalarning barcha noyob juft taqqoslamalari uchun normallashtirilgan d_H ning o'rtacha qiymati $\bar{x}$ quyidagi formula bilan hisoblanadi:

$$\bar{x} = \frac{\sum_{i=1}^k M_i}{k} \quad (3)$$

bu yerda,

M_i - turli namunalarning i -noyob juftligi uchun funksiya qiymati (yoki barcha diagonal bo'lmagan elementlar yig'indisining yarmi).

3-formulaga ko'ra, quyidagini qiymatlarni olamiz:

$$\bar{x}_{dH} = 4,8 / 10 = 0,48 \text{ yoki namunalar orasidagi farq qiymati } 48\% (0,48 \times 100).$$

Genetik tozalikni aniqlash ko'rsatkichlariga asoslanib, "A" navining o'xshashligi yoki tozalik darajasi uning urug'lik partiyasi doirasida mos ravishda 52% $((1 - 0,48) \times 100)$ ni tashkil etadi.

Levenshteyn masofasi (d_L) - turli uzunlikdagi qatorlarni taqqoslash imkonini beruvchi funksiya. U bir belgidan iborat operatsiyalarning minimal soni sifatida tavsiflanib, bir qatorni boshqasiga aylantirish uchun zarur bo'lgan bir belgini boshqasiga kiritish, olib tashlash va almashtirishlarni o'z ichiga oladi [5].

$A = a_1, a_2, \dots, a_m$ va $B = b_1, b_2, \dots, b_n$ ikki qator orasidagi d_L ni quyidagicha ifodalash mumkin:

$(m + 1) \times (n + 1)$ o'lchamli D matritsa berilgan bo'lsin,

Bu yerda,

m – birinchi qator uzunligi; n – ikkinchi qator uzunligi; $D(i, j)$ birinchi qatorning birinchi i belgilari va ikkinchi qatorning birinchi j belgilari orasidagi masofani ifodalaydi. Matrisani faollashtirish (inisializatsiyasi):

$$D(0, 0) = 0; D(i, 0) = i \text{ для } i = 1, 2, \dots, m; D(0, j) = j \text{ для } j = 1, 2, \dots, n$$

Rekursiv formula:

$$D(i, j) = \min \begin{cases} D(i-1, j) + 1 \\ D(i, j-1) + 1 \\ D(i-1, j-1) + \delta(a_i, b_j) \end{cases} \quad (4)$$

Bu yerda,

$$\delta(a_i, b_j) = 0, \Rightarrow a_i = b_j; \delta(a_i, b_j) = 1, \Rightarrow a_i \neq b_j$$

Aytaylik, «B» navining genetik tozaligini oltita SSR-markerlar yordamida baholashda beshta namunaga ($B_1 - B_5$) teng bo'lgan tanlanmada noaniq natijalarga

(N) ega bo'lgan pozitsiyalar (allel mavjudligi yoki yo'qligini aniqlashning imkoni yo'qligi) aniqlandi (3-jadval). Bunday vaziyatda, qatorlar har xil uzunlikka ega bo'lganligi sababli, d_H qiymatini hisoblash imkonsizdir.

3-jadval.

«B» navining allel profili

	SSR – marker					
Namuna	1	2	3	4	5	6
B ₁	1	1	0	1	1	N
B ₂	1	0	1	1	0	0
B ₃	0	1	1	0	1	N
B ₄	1	0	1	1	1	N
B ₅	1	1	0	1	1	1

4-formuladan foydalanib, Levenshteyn oralig'ini hisoblash uchun quyidagi argumentlarni kiritamiz: ixtiyoriy ikkita bir xil belgilarning mos kelishi (shu jumladan N va N) = 0; allellar o'rtasidagi almashtirish (0) va (1) = 1; (N) belgini ma'lum bir allelga almashtirish (0 yoki 1) = 2 (ya'ni (N) belgini olib tashlash va qatorlarning bir biriga o'xshashligi uchun (0) yoki (1) ni kiritish kerak). Namunalar bir-biri bilan o'zaro taqqoslanganda quyidagi d_L taqsimotiga ega bo'lamiz (4-jadval, a). Keyinchalik, «B» navining genetik tozaligini aniqlash uchun olingan d_L metrik qiymatlarini qator uzunligi bo'yicha normallashtiramiz - $d_L/\max(m, n)$ (4-jadval, b).

4-jadval

«B» nav namunalarini taqqoslashda d_L ko'rsatkichining taqsimlanishi

a)	B ₁	B ₂	B ₃	B ₄	B ₅	b)	B ₁	B ₂	B ₃	B ₄	B ₅
B ₁	0	5	3	2	2	B ₁	0,0	0,8	0,5	0,3	0,3
B ₂	5	0	6	3	4	B ₂	0,8	0,0	1,0	0,5	0,7
B ₃	3	6	0	3	5	B ₃	0,5	1,0	0,0	0,5	0,8
B ₄	2	3	3	0	4	B ₄	0,3	0,5	0,5	0,0	0,7

B₅	2	4	5	4	0	B₅	0,3	0,7	0,8	0,7	0,0
----------------------	---	---	---	---	---	----------------------	-----	-----	-----	-----	-----

2- va 3-formulalardan foydalangan holda namunalarning barcha noyob juft taqqoslash namunalari uchun normallashtirilgan d_L ning o'rtacha qiymatini hisoblaymiz:

$$\bar{x}_{dL} = 6,1 / 10 = 0,61 \text{ yoki namunalar orasida tafovut } 61\% (0,61 \times 100).$$

O'z navbatida, «B» navining genetik tozalik darajasi uning urug'lik partiyasi ichida 39% $((1 - 0,61) \times 100)$ ni tashkil etadi.

XULOSA

Navning genetik tozalil darajasini hisoblash uchun ko'rsatkichlarni tahlil qilish shuni ko'rsatdiki, allel profillarining katta ma'lumotlar to'plamini hisoblash uchun eng maqbul tanlov taqqoslanayotgan qatorlarning bir xil uzunligi sharoitida Xemming oralig'idan foydalanishdir. Aksincha, Levenshteyn oralig'i o'zgaruvchan uzunlikdagi qatorlar bilan ishlashga imkon beradi, ammo hisoblashning o'ta murakkabligi bilan ajralib turadi. Keyingi tadqiqotlarning istiqbolli yo'nalishi bu, o'zgaruvchan uzunlikdagi qatorlar uchun ma'lumotlarni qayta ishlashning yuqori tezligini ta'minlaydigan metrikalarni tahlil qilishdir. Shuningdek, allel profillarining haqiqiy va sezilarli hajmli ma'lumotlar massivlari bilan ishlashda o'rganilayotgan ko'rsatkichlarning xususiyatlarini o'rganish muhim ahamiyatga ega bo'lib, bu navning genetik tozaligi tahlilining samaradorligi va aniqligini oshirishga imkon beradi.

Adabiyotlar ro'yxati

1. Genetic Purity Testing of F1 Hybrids of Cotton Using DNA Markers / G. M. Puneeth, S. Deo, L. Arya, M. K. Rana // Journal of Scientific Research and Reports. 2024. Vol. 30, No. 9. P. 537–542. DOI: 10.9734/jsrr/2024/v30i92380.
2. Molecular Techniques for Testing Genetic Purity and Seed Health / V. Santhy, N. Sandra, K. Ravishankar, B. Chidambara // Seed Science and Technology: Springer, 2023. P. 365–389. DOI: 10.1007/978-981-19-5888-5_15.
3. Hamming, R. W. Error detecting and error correcting codes // The Bell System Technical Journal. 1950. Vol. XXIX, No. 2. P. 147–160.

4. Пономарева, Н. С., Реброва, Г. Н., Колина, Е. А. Применение расстояний редактирования при биоинформационном анализе геномов для задач оценки состояния репродуктивной системы // Фундаментальные исследования. 2015. № 7-4. С. 774-777.
5. Левенштейн, В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии наук СССР. 1963. № 4. С. 845-848.